COMPARING CHARGES, LENGTH OF STAY, AND MORTALITY FOR A PROVIDER TO AN INTERNAL OR EXTERNAL NORM

INTRODUCTION

All Patient Refined DRGs (APR-DRGs) are a patient classification system that categorizes patients based on their severity of illness and risk of mortality. In APR-DRGs there are 314 base APR-DRGs, each of which is subdivided into four subclasses for a total of 1256 unique patient categories. There are extensive normative data available for APR-DRGs. These normative data include average values of resource variables such as length of stay and charges and the percent occurrence of outcomes such as mortality. The normative data available include norms by geographic region, by different types of hospitals and by payors.

Based on APR-DRGs, individual providers can compare their resource use and outcomes to the normative data to determine if their performance differs significantly from the normative data. Observed differences between a provider's performance and the norm can represent a true difference in performance or can be caused by random variation. Statistical methods can be used to determine which differences in resource use or outcomes are true differences and which may be the result of random variation.

The statistical methods give the probability that an observed difference in performance between the provider and the norm is due to random variation. A difference in performance between provider and norm is considered "significant" if this probability is small. A difference is considered significant at the 0.05 level if the probability that the observed difference is due to random variation is five percent or less (i.e., less than one chance in twenty). Significance at the 0.01 level means that this probability is one percent or less.

Three interrelated factors determine whether a difference in performance is significant: the number of observations, the magnitude of the observed difference in performance, and the variability in performance of the hospital and of the norm. A small number of patients, a small observed difference in performance, or high variability within either the provider or the norm (i.e., high standard deviation) increase the probability that the observed difference is due to chance and does not represent a true difference. Conversely, a large number of patients, a large observed difference between provider and norm, or low variability within both hospital and norm make it more likely that the difference was not due to chance and does represent a true difference. An observed difference of the same magnitude may be significant in one comparison and not in another. For example, a half-day difference in average length of stay for normal delivery, with a large number of patients and low variability, is unlikely to be due to

random variation and is, therefore, considered significant. However, for transplant surgery with few patients and high variability of length of stay, a half day difference is more likely to be due to random variation and not be considered significant. The conclusion that a difference is significant indicates that the hospital and the norm have a true difference in performance, which is likely to be repeated in future data. For normal delivery, an observed difference of a half day may be enough evidence to reach this conclusion but the same observed difference may yield only weak, inconclusive evidence for transplants.

There are several possible reasons why a difference may not be significant. There may be no true difference, and thus, no significant difference in performance is found. Alternatively, there may be too few observations or too much variability, or both, so that even a true difference can not be detected. Thus, a difference which is not found to be significant does not necessarily mean that there is no true difference in performance. It may simply mean that there were too few patients or too much variability to conclude that there is a true difference.

The comparison of a provider's performance to a norm requires the use of several distinct statistical methods. Resource variables such as length of stay and charges are continuous variables that, in general, are lognormally distributed, while outcome variables are binary variables that indicate the occurrence or non occurrence of an event such as death. Different statistical methods are required for continuous and binary variables. Comparisons can be performed for data from a single APR-DRG and subclass or can be performed for data pooled across multiple APR-DRGs and subclasses. Different statistical methods are required for the comparisons of data within a single APR-DRG and subclass versus the comparison of data pooled across multiple APR-DRGs and subclasses. The norm to which the provider is compared may contain the data from the provider. If the norm contains data from the provider, then before any statistical tests are performed the provider's data must be backed out of the norm.

This document describes the various statistical methods used to determine the statistical significance of the difference between the average value of a resource variable for a provider and the average value derived from normative data and the statistical significance of the difference between the rate of occurrence of an outcome variable for a provider and the rate of occurrence derived from normative data. The statistical methods for continuous variables (e.g., length of stay and charges) are presented first followed by the statistical methods for binary variables (e.g., mortality). Within each of these sections, the statistical method for comparison of data from a single APR-DRG and subclass is presented first followed by the statistical method for the comparison of data pooled across multiple APR-DRGs and subclasses. Finally, the method used to back out a provider's data from a norm is described. A discussion of the statistical

methods described in this document can be found in Cohen, 1977 and Kramer and Thiemann, 1987.

CONTINUOUS VARIABLES

Resource variables tend to have lognormal distributions, meaning that, within a homogeneous subclass of patients, the logarithm of these variables (to base e or base 10) has roughly a bell-shaped, normal distribution. Since resource variables tend to have a lognormal distribution, one might assume that a logarithmic transformation was needed to ensure the validity of the statistical comparison. However, a logarithmic transformation is not necessary for comparing the average values of resource variables. This is true because averages have a normal distribution even if the raw data do not. Furthermore, the difference of two averages, under the null hypothesis that the two averages are equal, has a symmetric distribution, regardless of the asymmetry of the distribution of the underlying observations (Chung, 1968, p. 137). The statistical test that compares the difference in two averages will behave correctly (maintain correct type I error) when the two averages being compared are based on the non logarithmic scale. Thus, the comparison of resource variables will be based on the average value of the "raw" observations, not of their logarithms.

There remains an important issue, namely that an average is not the ideal measure of central tendency for data with a lognormal distribution, especially when a large coefficient of variation indicates a strongly right-skewed distribution (Cohen, Whitten, 1988, Chapter 4). In such a case, the average may substantially exceed the median with the average more sensitive than the median to occasional very large data values. The statistical methods used assume that occasional, aberrant, extreme observations have been removed from both ends of the data, very small and very large. High and low length of stay trim points have been developed for each APR-DRG and subclass. These length of stay trim points are used to eliminate aberrant, extreme observations from the comparison of average values of resource variables. After eliminating aberrant, extreme observations, the comparison of averages is done in the original, non logarithmic scale.

One more technical condition refers to large numbers of values of zero. For example, many patients may not have an ICU charge. Thus, the distribution of ICU charges will have a “bump” at zero. Such distributions do not represent continuous data and the statistical methods described for continuous data do not apply to variables with a large number of tied values, such as many values equal to zero.

<u>COMPARISON WITHIN A SINGLE APR-DRG AND SUBCLASS</u>

OUTLINE OF APPROACH: The analysis begins by a comparison of variances between the provider and the norm. Student's t-tests are then used to compare the provider average with the norm average, with the details varying according to the situation.

NOTATION: There are two groups of data (e.g., a provider and the norm), with m observations in the provider and n in the norm. The observations are

$$x_{11}, ..., x_{1m} \text{ (provider) and } x_{21}, ..., x_{2n} \text{ (norm)}$$

Let $\underline{x}_1$ and $\underline{x}_2$ be the two averages, $\Sigma x_{1j}/m$ and $\Sigma x_{2j}/n$, respectively. The two sample variances are

$$s_1^2 = \Sigma_j(x_{1j}-\underline{x}_1)^2/(m-1) \text{ and } s_2^2 = \Sigma_j(x_{2j}-\underline{x}_2)^2/(n-1)$$

COMPUTATIONAL STEPS

If either m or n is five or less the method being used to test the significance of the difference in averages is unreliable and no test of significance is performed.

Step 1. CHECK IF THE VARIANCES ARE ROUGHLY EQUAL.
Compute the ratio

$$F = \text{larger of } (s_1^2, s_2^2)/\text{smaller of } (s_1^2, s_2^2)$$

Under the assumption that the true variances are indeed equal, this ratio has an F-distribution with (m-1, n-1) degrees-of-freedom (Madansky, 1988, p.59). If s_1^2 or s_2^2 is zero the method being used to test the significance of the difference in averages is unreliable and no test of significance is performed. The variances are considered equal if the above ratio has a value that is less than 4 and is also below the critical value of the F distribution for a one tail test at a 0.001 level of significance.

Step 2. EQUAL VARIANCES FOUND IN STEP 1.
Compute the two-sample, equal-variance Student's t-statistic. This is given by

$$t^* = (\underline{x}_1 - \underline{x}_2)/\sqrt{(A*B)} \text{ where}$$

$$A = (1/m) + (1/n) \text{ and } B = \{(m-1)s_1^2 + (n-1)s_2^2\}/(m+n-2)$$

Under the null hypothesis that the true mean of the provider and norm are equal, t* has a Student's t-distribution with m+n-2 degrees-of-freedom (Hogg, Craig, 1978, p. 264).

The averages are considered to be significantly different if t* has a value outside of the critical values of the t distribution for a two tail test at a specified level of significance.

Step 3. UNEQUAL VARIANCES FOUND IN STEP 1.

The two-sample, unequal variance t-statistic is given by

$$t^* = (\underline{x}_1 - \underline{x}_2)/\sqrt{(s_1^2/m + s_2^2/n)}$$

The degrees-of-freedom (Satterthwaite, 1946) are

$$df = (s_1^2/m + s_2^2/n)^2/(s_1^4/m^2(m-1) + s_2^4/n^2(n-1))$$

Under the null hypothesis that the true mean of the provider and norm are equal, t* has a student t distribution with df degrees of freedom. The averages are considered to be significantly different if t* has a value outside of the critical values of the t distribution for a two tail test at a specified level of significance.

COMPARISON OF A CONTINUOUS VARIABLE FOR DATA POOLED ACROSS APR-DRGS AND SUBCLASSES

OUTLINE OF APPROACH: The comparison is based on the principle of stratification. The average of the provider and the norm are compared separately within each APR-DRG and subclass. A weighted average of the difference in averages, one per APR-DRG and subclass is formed. If the averages for the provider and norm are equal, the weighted average difference should be close to zero.

NOTATION:

- i = group (1 = provider and 2 = norm)
- j = APR-DRG and subclass combination
- m = mth patient
- K = number of APR-DRG and subclass combinations
- x_{ijm} = value of the resource variable for the mth patient in APR-DRG and subclass combination j in group i
- $n(i,j)$ = number of patients in APR-DRG and subclass combination j in group i
- $\underline{x}_{ij}$ = average value of the resource variable for patients in APR-DRG and subclass combination j in group i
- s^2_{ij} = variance of the resource variable for patients in APR-DRG and subclass combination j in group i

d_j = difference between the average in the two groups for APR-DRG and subclass combination j

v_j = variance of the difference between the average in the two groups for APR-DRG and subclass combination j

Provider

In subclass 1, there are n(1,1) observations: $x_{111}, x_{112}, \ldots x_{11n(1,1)}$, average $\underline{x}_{11}$, variance s_{11}^2
In subclass 2, there are n(1,2) observations: $x_{121}, x_{122}, \ldots x_{12n(1,2)}$, average $\underline{x}_{12}$, variance s_{12}^2
...
In subclass k, there are n(1,k) observations: $x_{1k1}, x_{1k2}, \ldots x_{1kn(1,k)}$, average $\underline{x}_{1k}$, variance s_{1k}^2

Norm

In subclass 1, there are n(2,1) observations: $x_{211}, x_{212}, \ldots x_{21n(2,1)}$, average $\underline{x}_{21}$, variance s_{21}^2
In subclass 2, there are n(2,2) observations: $x_{221}, x_{222}, \ldots x_{22n(2,2)}$, average $\underline{x}_{22}$, variance s_{22}^2
...
In subclass k, there are n(2,k) observations: $x_{2k1}, x_{2k2}, \ldots x_{2kn(2,k)}$, average $\underline{x}_{2k}$, variance s_{2k}^2

COMPUTATIONS: For each APR-DRG and subclass the difference in the averages is formed for all values of j from 1 to K.

$$d_j=\underline{x}_{1j}-\underline{x}_{2j} \text{ with variance } v_j=s_{1j}^2/n(1,j)+s_{2j}^2/n(2,j)$$

If there is only one patient for the provider (i.e., n(1,j) = 1) then the variance of the provider is estimated as

$$s^2_{1j} = (\underline{x}_{1j}(s_{2j}/\underline{x}_{2j}))^2$$

The weighted average difference between the provider and the norm is

$$D = \Sigma_j(d_j/w_j) \,/\, \Sigma_j(1/w_j) \text{ where } w_j = 1/n(1,j) + 1/n(2,j)$$

Under the null hypothesis that, within each APR-DRG and subclass combination, the true means for the provider and the norm are equal, the quantity D has a normal distribution with mean 0 and variance

$$V = \sum_j v_j/w_j^2/(\sum_j(1/w_j))^2$$

Thus, the test of equal means is based on $z = D/\sqrt{V}$ which is compared to a standard normal distribution with mean 0 and variance 1. The averages are considered to be significantly different if D has a value outside of the critical values of the standard normal distribution for a two tail test at a specified level of significance.

In the computatation of the weighted average difference between the provider and the norm (i.e., D), w_j is based only on the sample sizes. Alternatively, w_j could be based on the variance of d_j

$$v_j=s_{1j}^2/n(1,j)+s_{2j}^2/n(2,j)$$

A test statistic computed using weights based on the variance of d_j (i.e., v_j) would produce the highest chance of yielding a true positive finding (i.e., maximum power). However, a test statistic computed based on v_j is dominated by small values of s^2_{1j} even when n(1,j) is small. This is particularly true when the sample sizes in the norm are large so that s^2_{2j}/n (2,j) is very small. Thus, although a test statistic computed based on v_j would have the maximum power, the test statistic based on w_j is used since it yields a more robust estimate (i.e., less sensitive to small values of s_{1j}^2).

Individual providers are unlikely to treat patients in every APR-DRG and subclass. For example, most hospitals do not perform heart transplant surgery. Thus, there is the expectation that the data for the provider will contain no patients in some APR-DRGs and subclasses. In general, norms are derived either from large databases which are expected to include patients from virtually all APR-DRGs or from specific patient populations (e.g., pediatrics or Medicare) in which data are only present for a subset of the APR-DRGs and subclasses. In any comparison of a provider and a norm, those APR-DRGs and subclasses in which the provider has no patients (i.e., n(1,j) is zero) are not included in the comparison. In addition, APR-DRGs and subclasses in which there are either no patients or only one patient in the norm (i.e., n(2,j) is zero or one) are also excluded from the comparison. Although the statistical test being used is not very sensitive to the number of APR-DRGs and subclasses being excluded, the exclusion of a substantial number of APR-DRGs and subclasses due to the presence of either no patients or only one patient in the norm raises questions concerning the validity of the norm and the generalizability of the results of the comparison. For example, if a provider's Medicare patients were compared to a pediatric norm many of the APR-DRGs and subclasses present in the provider's data would not be present in the norm. Thus, such a comparison is not meaningful and the results of any test of significance of the difference in performance between the provider and norm would not be valid. In order to protect against the inappropriate application of a test of significance, if ten percent or more of the patients in the subset of the provider's data being compared are excluded from the comparison because there is either no patients or only one patient in the norm, no test of significance of the difference in averages between the provider and the norm is performed. Since the norm is, in general, expected to contain virtually all the APR-DRGs and subclasses in the subset of the provider's data being compared, the exclusion of ten percent of the provider's patients from the comparison is considered an

indication that there is an incompatibility between the provider's data and the norm and, therefore, any comparison would not be meaningful.

If the number of patients pooled across APR-DRGs and subclasses in the provider or the norm is five or less the method used is unreliable and no test of significance is performed.

REFERENCES

Agresti, A, *Categorical Data Analysis*, Wiley, New York, 1990.

Chung, K, *A Course in Probability Theory*, Harcourt, Brace, New York, 1968.

Cohen, A, Whitten, B, *Parameter Estimation in Realibility and Life Span Models*, Marcel Dekker, New York, 1988.

Cohen, J, *Statistical Power Analysis for the behavioral Sciences*, Academic Press, New York, 1977.

Hogg, R, Craig A, *Introduction to Mathematical Statistics, Fourth Fourth Edition*, Macmillan, New York, 1978.

Kraemer, H, Thiemann, S, *How Many Subjects? Statistical Power Analysis in Research*. Sage Publications, Newbury Park, California, 1987.

Madansky, A, *Prescriptions for Working Statisticians*, Springer-Verlag, New York, 1959

Mantel, N, Haenszel, W, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease", *Journal of the National Cancer Institute*, Vol. 22, 1959, p. 719-748.

Satterwaithe, F, "An Approximate Distribution of Estimates of Variance Components", *Biometrics Bulletin*, Vol. 2, 1946, p. 110-114.

